### Abstract
- Max-margin learning of latent-variable sequence predictors
- Function class: Observable Operator Models
- Contributions:
  - Max-margin completion of a Hankel matrix
  - Spectral regularisation for structured prediction
- Learning formulated as convex optimisation

### Sequence Tagging
- Task: map input sequences to output sequences
  - **hippopotamus**
  - **hlp x patm x s**
- Task Loss \( f(x, y) \): Hamming distance
- Formulation using a scoring function
  \[ F(X \times Y)^T \rightarrow \mathbb{R} \]
  \[ \hat{g}(x) = \arg \max_{y \in Y} F(x, y) \]
- Max-margin Structured Prediction: given \( m \) training examples and a class of functions \( \mathcal{H} \) solve
  \[ \arg \min_{H \in \mathcal{H}} \sum_{i=1}^{m} L(x_i, y_i; F) + \tau \| H \| \]
  - **non-convex**

### Usual Function Classes
- Factorized Linear Models of order \( k \) (e.g. CRF)
  \[ F(x, y) = \sum_{t=1}^{T} w \cdot \Phi(x, y_{t-1-k}) \]
  - **tractable but very dependent on \( \Phi \)**
- Latent-variable Models of order \( k \) (e.g. latent SVM, HCRF)
  \[ S(x, y, h) = \sum_{t=1}^{T} w \cdot \Phi(x, y_{t-1-k}, h_{t-k}) \]
  \[ F(x, y) = \sum_{h} \exp(S(x, y, h)) \]
  - **intractable and learning**

### Definition
- \( A = \{X, Y, n, \alpha, (A_{ab}), \beta \} \)
- Alphabets: input \( X \), output \( Y \)
- \( n \) states (i.e., hidden dimensions)
- Initial weights \( \alpha \in \mathbb{R}^n \)
- Final weights \( \beta \in \mathbb{R}^n \)
- Operator for each bi-symbol \( A_{ab} \in \mathbb{R}^{n \times n} \)

### Prediction with IO-OOM
- Scoring function using \( h \)
  \[ S(x, y, h) = \alpha(h_0) \prod_{t=1}^{T} A_{a_{x_t} y_t}(h_{t-1}, h_t) \beta(h_T) \]
- Scoring function marginalizing \( h \)
  \[ F(x, y) = \sum_{h} S(x, y, h) = \alpha^T A_{a_{x_1}} \cdots A_{a_{x_T}} \beta \]
- Global piecewise prediction of order \( k \)
  \[ \hat{y}_k(x) = \arg \max_{y} \sum_{t=k+1}^{T} F(x_{t-k}, y_{t-k}) \]

### The Hankel Matrix of IO-OOM
- \( \mathcal{X} = \{a, b, c\} \)
  \[ \mathcal{Y} = \{\triangle, \nabla\} \]

### Convex Optimization
- Objective using Hankel and rank
  \[ \arg \min_{H \in \mathbb{R}^{n \times n}} \sum_{i=1}^{m} L(x_i, y_i; H) + \tau \| H \| \]
  - **non-convex**
- Objective using Hankel and nuclear-norm
  \[ H_S \in \arg \min_{H \in \mathbb{R}^{n \times n}} \min_{k} \sum_{i=1}^{m} L(x_i, y_i; H) + \tau \| H \| \]
  - **convex**
- We use Forward Backward Splitting (FOBOS) (Duchi & Singer 2011), based on gradient steps and proximal operators.

### Experiments
- Phonetic transcription, “Nettalk” data (\( |A| = 26, |Y| = 51 \)) (Sejnowski & Rosenberg, 1987)
- Methods compared:
  - IO-OOM Max-margin
  - Spectral regularisation
  - L^2 regularisation
  - no regularisation
  - Factorized linear model (CRF, avg. perceptron)
  - IO-HMM trained directly with spectral method

### Acknowledgements
This work was supported by projects XLike (FP7-288342), ERA-Net CHISTERA VISETN, TACARDI (TIN2012-38523-C02-00) and by the IFSC Centers of Excellence (grant 1789/11). X. Carreras was supported by the Ramón y Cajal program of the Spanish Government (RYC-2008-02223). B. Balle received support from NSERC and the James McGill Research Fund. Xerox Research Centre Europe (XRCE) provided travel support for A. Quattoni and X. Carreras.